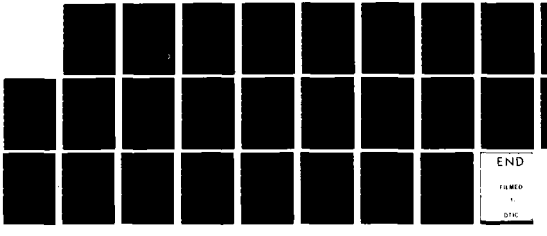
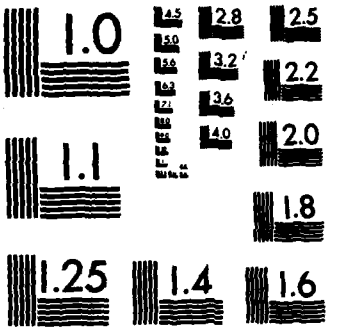
AD-A124 787 NEAR INFRARED REFLECTANCE ANALYSIS BY GAUSS-JORDAN LINEAR ALGEBRA(U) INDIANA UNIV AT BLOOMINGTON DEPT OF CHEMISTRY D E HONIGS ET AL. 10 FEB 83 1/1

UNCLASSIFIED INDU/DC/GMH/TR-83-53 N00014-76-C-0838 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: INDU/DC/GMH/TR-83-53 — 2. GOVT ACCESSION NO.: AD-A124787 — 3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Near Infrared Reflectance Analysis by Gauss-Jordan Linear Algebra	5. TYPE OF REPORT & PERIOD COVERED: Interim Technical Report
	6. PERFORMING ORG. REPORT NUMBER: 62
7. AUTHOR(s): D.E. Honigs, J.M. Freelin, and G.M. Hieftje	8. CONTRACT OR GRANT NUMBER(s): N14-76-C-0838
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Chemistry, Indiana University, Bloomington, IN 47405	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 051-622
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, D.C.	12. REPORT DATE: 10 February 1983
	13. NUMBER OF PAGES: 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Prepared for publication in APPLIED SPECTROSCOPY

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

near infrared reflectance
signal processing

DTIC ELECTE FEB 23 1983 S D E

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Near-infrared reflectance analysis (NIRA) is an analytical technique that uses the near-infrared diffuse reflectance of a sample at several discrete wavelengths to predict the concentration of one or more of the chemical species in that sample. However, because near-infrared bands from solid samples are both abundant and broad, the reflectance at a given wavelength usually contains contributions from several sample components, requiring extensive calculations on overlapped bands. In the present study, these calculations have been performed

AD A124787

DTIC FILE COPY

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

83 02 0[illegible] 002

20. Abstract (continued)

using an approach similar to that employed in multi-component spectrophotometry, but with Gauss-Jordan linear algebra serving as the computational vehicle. Using this approach, correlations for percent protein in wheat flour and percent benzene in hydrocarbons have been obtained and are evaluated. The advantages of a linear-algebra approach over the common one employing stepwise regression are explored.

OFFICE OF NAVAL RESEARCH

Contract N14-76-C-0838

Task No. NR 051-622

NEAR INFRARED REFLECTANCE ANALYSIS BY GAUSS-JORDAN LINEAR ALGEBRA

by

D. E. Honigs, J. M. Freelin, G. M. Hieftje, and T. B. Hirschfeld

Prepared for Publication

in

APPLIED SPECTROSCOPY

Indiana University
Department of Chemistry
Bloomington, Indiana 47405

February 10, 1983

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

DTIC COPY INSPECTED 2

Reproduction in whole or in part is permitted for any purpose of the United States Government

This document has been approved for public release and sale; its distribution is unlimited

INTRODUCTION

A. Overview. The application of near-infrared reflectance analysis (NIRA) as an analytical technique has been concentrated mainly in the agricultural area where it originated.[1-4] These agricultural applications are characterized by the need to determine a limited number of constituents in a very large number of individual but similar samples. In contrast to this situation, samples encountered in most industrial analytical laboratories are widely varied in kind and the number of very similar samples is limited.

If NIRA is to be applied broadly to industrial chemical analysis, it must be modified to sharply reduce the developmental effort needed to set up an individual method. At present, the establishment of a NIRA procedure requires the assembly of a fairly large set of standard samples where composition has already been established by a reference method or methods. The reflectance of the samples must then be measured at a substantial number of points in the near-infrared spectral region, and the resulting data subjected to a multilinear regression algorithm. This algorithm then generates the choice of analytical wavelengths and yields a "correlation equation" which relates concentration of desired constituents to reflectance at various near-infrared wavelengths. This latter calculation step is very demanding of computer hardware and processing time, particularly for high-performance NIRA instruments that cover a wide range of wavelengths, a probable requirement for industrial analysis. In fact, these computational requirements are so demanding that they have often forced shortcuts in methods development and optimization, and have limited the performance of the NIRA system. This limitation is manifested by incomplete optimization of the analytical wavelength set, whereby either too few wavelengths are examined, abridged wave-

length selection criteria are used, or too few or incorrect wavelengths are selected for the NIRA procedure.

This limitation also encourages incomplete testing and evaluation of the developed NIRA method, which leads to the widespread use of poorly understood algorithms and the retarded development of improved ones. This lack of understanding and the workload of existing methods have often been great enough to discourage people from examining the NIRA approach, and have acted as a brake on its wider acceptance and application.

In the present paper, an algorithm is described and evaluated for substantially accelerating the wavelength and calibration coefficient selection process of NIRA. This algorithm is used to find "correlation equations" for protein in wheat and benzene in a hydrocarbon mixture. Bias-corrected standard errors of prediction obtained with the new algorithm reached 0.26 percent protein in wheat and 1.01 percent benzene by volume. Comparisons of the algorithm with several others based on regression show improvements in computation time ranging from a few percent to as much as 200-fold. It is also discussed how the novel method might prove advantageous in the reduction of overfitting and in the improvement of NIRA accuracy.

B. Calibration Procedures in NIRA. The general pattern for establishing a NIRA calibration is described in a review article by Watson,[5] and will be briefly summarized here for clarity. The first step in establishing a NIRA calibration is to obtain a sample set in which the desired characteristic or sample constituent has been previously determined by a reference chemical or spectroscopic technique. An example of such a set would be wheat samples whose protein content had been established by Kjeldahl determinations. The sample set is randomly divided into two subsets, one for solving the regression

procedure (training) and one for testing the regression (prediction). Next, the near-infrared diffuse-reflectance spectrum for each sample is obtained. The spectra from the training set are then analyzed by some form of multiple linear regression. Typically, -log reflectance values (R) are regressed against the chemically determined concentrations to identify a group of wavelengths at which R best predicts the desired constituent in the training set. A number of alternative linear regression techniques are currently available to establish the NIRA calibration. These techniques include (but are not limited to) stepwise, all possible combinations, all possible pairs stepwise, and all possible triplets stepwise.

Stepwise regression is well known in statistical applications.[6] In its most general form a stepwise regression algorithm calculates the linear regression between two sets of variables and establishes a statistical confidence level to their degree of coherence. It then adds new values to or deletes old ones from one of the sets in an attempt to improve the coherence; coherence is usually expressed in terms of a correlation coefficient. Procedures involving the addition or deletion of values are called forward stepwise and backward stepwise regression, respectively. In its application to NIRA, forward stepwise regression involves the addition of R values at new wavelengths and suffers from the problem that the newly added wavelength is often not the best wavelength to add. Moreover, background interferences can cause omission of an important wavelength. Backward stepwise regression in NIRA requires that the total number of wavelengths that are employed be small enough that the regression containing all wavelengths can be calculated in a reasonable time, as the starting point for the backwards stepping. This requirement is usually inconsistent with the amount of data generated by a spectral scanning instrument.

The "all-possible-combinations" regression[6] improves upon the forward stepwise approach, in that background interferences do not bias the selection of wavelengths. The drawback to the all-possible-combinations technique is the enormous number of calculations it requires. This number is equal to 2^m, where m is the total number of wavelengths being employed, restricting this approach to those applications with the smallest data sets.

In an effort to combine the advantages of the stepwise and all-possible-combinations methods, several hybrid techniques such as "all-possible-pairs stepwise" and "all-possible-triplets stepwise" have been developed.[7] These techniques begin with all possible pairs or triplets of wavelengths, respectively, and proceed by means of a forward stepwise regression. In this way, the best pair or triplet of wavelengths can not be hidden by background interferences, yet the number of required calculations is much less than in the all-possible-combinations method. To ensure self-consistency, one of the wavelengths earlier adopted in the calibration is dropped and the best wavelength to add is then determined by stepwise regression. If the calibration is self-consistent, this new wavelength is the same as the one just deleted. If not, the new wavelength is retained, a different one deleted, and the process repeated until the wavelength which is deleted is subsequently restored by the regression process.

After using any of these regression techniques one obtains a calibration of the form:

$$C = B_0 + B_1\underline{R}_1 + B_2\underline{R}_2 + \ldots + B_j\underline{R}_j \qquad (1)$$

where $B_0 \ldots B_j$ are the coefficients of intercept and partial slopes from

the regression equation, $\underline{R}_j$ is -log (reflectance) of the sample at the jth wavelength and C is the concentration of the desired species in the sample.

Once the B_0 through B_j coefficients are determined, the standard deviation between the actual and predicted concentrations for the training set (corrected for the statistical degrees of freedom) is computed and called the "standard error of estimation" (SEE). The mathematical definition of SEE is given in Eq. (2).

$$SEE = \left[(N_S - 1 - N_W)^{-1} \sum_{i=1}^{N_S} e_i^2 \right]^{1/2} \tag{2}$$

where N_S is the number of samples in the training set, N_W is the number of wavelengths kept and e_i is the difference between the true component concentration and the value predicted by Eq. 1 for the i^{th} sample.

Next, the deduced regression equation (Eq. 1) is used to calculate the concentration of the desired constituent in each of the samples in the prediction set. From these computed concentrations and those known from the earlier independent chemical analysis (e.g. Kjeldahl determination), another standard deviation is determined, termed the "standard error of prediction" (SEP). The definition of SEP is given in Eq. (3).

$$SEP = \left[(N_S' - 1)^{-1} \sum_{i=1}^{N_S'} e_i^2 \right]^{1/2} \tag{3}$$

where N_S' is the number of samples in the prediction set.

The value of SEP is typically used as a measure of the performance of Eq. 1; however, a bias-corrected SEP better estimates how well the calibration will perform in the field, where routine comparisons between NIRA results

and results from the reference chemical method are periodically used to adjust the long-term drift of the NIRA spectrophotometer. This bias-corrected SEP is given by the equation:

$$\text{SEP (biased)} = \left[(N_S' - 1)^{-1} \sum_{i=1}^{N_S'} (e_i - \text{Bias})^2\right]^{1/2} \tag{4}$$

where

$$\text{Bias} = (N_S')^{-1} \sum_{i=1}^{N_S'} e_i \tag{5}$$

C. Row-reduction. Because NIRA is similar to multi-component uv-visible spectrophotometry, it would be very useful to transfer the knowledge and technology of this latter field to NIRA. Unfortunately, this transfer is not straightforward. A NIRA spectrum contains virtually no peaks attributable to a single species, so individual "absorptivities" cannot be measured and background corrections are very complex. In fact, it was this very complexity that led to the introduction of regression techniques. Unfortunately, multilinear regression techniques are very easily overfitted and can be very slow.

In an attempt to reduce the overfitting of multilinear regression and shorten computation time, a simplifying assumption has been made in the present study. Specifically, if the errors in the reference chemical method and in the measured diffuse reflectance spectrum are small, a simple linear-algebra solution of j unknowns with j equations will give a good first approximation to a multi-linear regression. To test the assumption, the Gauss-Jordan reduction[8] method for treating linear equations was used to

solve Eq. (1) for several NIRA sample sets. The authors have elected to call this particular application of Gauss-Jordan reduction "row-reduction".

I. THEORY

Gauss-Jordan reduction is a general approach to solving a system of n equations for n unknowns. A full description of the mathematics involved can be found in Reference 8. Briefly, to solve for a single variable in a system of equations such as the one shown in Eq. (6) [which can be re-written in matrix form as Eq. (7)], each equation can be multiplied by some constant and then subtracted from another equation. For example, to solve Eqs. (6) and (7) for the variable x, the second row can be multiplied by -1/2 and the third row multiplied by -3/2. These operations transform Eq. (7) into Eq. (8).

$$\begin{aligned} 3x + 2y + 3z &= 16 \\ 6x + 2y + 8z &= 28 \\ 2x + 6y + 4z &= 26 \end{aligned} \tag{6}$$

$$\left[\begin{array}{ccc|c} 3 & 2 & 3 & 16 \\ 6 & 2 & 8 & 28 \\ 2 & 6 & 4 & 26 \end{array}\right] \tag{7}$$

$$\left[\begin{array}{ccc|c} 3 & 2 & 3 & 16 \\ -3 & -1 & -4 & -14 \\ -3 & -9 & -6 & -39 \end{array}\right] \tag{8}$$

When row 1 of Eq. (8) is added to rows 2 and 3, the resulting matrix is shown in Eq. (9).

$$\left[\begin{array}{ccc|c} 3 & 2 & 3 & 16 \\ 0 & 1 & -1 & 2 \\ 0 & -7 & -3 & -23 \end{array}\right] \tag{9}$$

This process can be continued to solve for y and z. When the matrix is entirely solved, it is in the form shown in Eq. (10), where I is the identity

$$\left[\begin{array}{c|c} & A_1 \\ I & A_2 \\ & \vdots \\ & A_p \end{array}\right] \tag{10}$$

matrix and A_p is the answer to the p^{th} variable in the equation. For Eq. (6), x is the first variable, y is the second and z is the third, and their solutions are found in A_1, A_2 and A_3, respectively.

The adaptation of Gauss-Jordan reduction to row-reduction NIRA is straightforward. The calibration of a NIRA sample set proceeds through the collection of spectra as described earlier. After the diffuse reflectance spectrum of each sample is obtained, the first j [where j is the number of terms in Eq. (1)] reflectance (R) values in the spectrum for each sample in the training set are placed in a matrix. It is important to recognize that the j reflectance values used in this matrix do not constitute the entire sample spectrum. Rather, they are merely the first j values of the

entire spectrum. The standard concentration of the sought-for species in each sample of the training set is also placed in the matrix; these concentrations correspond to the known values (right-hand side) of Eq. (4) and are termed the augmented portion of the matrix. The resulting matrix is shown pictorially in Fig. 1. Figure 1 is just the matrix form of the set of equations (like Eq. 1) resulting from several samples. The unknowns on the left-hand side of the matrix illustrated in Fig. 1 correspond to the B_1 through B_j values of Eq. (1).

To solve the matrix in Fig. 1 via Gauss-Jordan reduction, the matrix rows are rearranged to cause the largest reflectance (R) value in each column to lie on the diagonal. This rearrangement is called row interchanging. The matrix in Fig. 1 is then solved successively for each B_n term as described earlier; as a consequence, the remaining R values become orthogonal to those which were used to solve for the B terms. This behavior can be seen in Eq. (9); the first row of Eq. (9) is the only one which contains information about the unknown value x. When the matrix is completely solved and reduced to the form of Eq. (10), the first row is orthogonal to the rest of the matrix and contains information only about x. By means of row interchanging, the most mutually orthogonal samples are chosen to determine the B_n terms.

After the B_n values have been found, the solution is validated by comparing actual vs. predicted values for the training set and calculating a SEE and a correlation coefficient (r value). This r value is saved for comparison with later solutions.

Once the r value for the first matrix has been computed, the column corresponding to the wavelength with the largest B multiplier is dropped from the matrix and the R values for the next (j + 1) wavelength are put in its place. The computation and matrix solution are then repeated. After

every wavelength that was recorded in the original spectrum has proceeded through this computation, the entire process is repeated, using the final matrix as a starting point. All wavelengths are again stepped through the matrix solution procedure; after this second iteration, the combination which gave the best r value is recalled and used as the solution to Eq. (1).

In the wavelength-stepping procedure, the dropping of the column with the largest B value has an interesting effect. If reflectances at all wavelengths have roughly equivalent magnitudes, a reasonable assumption in the near infrared, the wavelength with the largest B value will contain the most information about the sought-for species. Because it is this "most important" wavelength that is dropped, the selection operation rapidly collects the most orthogonal wavelengths (those least correlated with the desired constituent and most correlated with background). This same selection criterion prevents the matrix from becoming "ill determined" and therefore subject to large roundoff error. An "ill determined" matrix typically contains very large positive and negative B values in pairs. Because the largest positive B value will be dropped by the selection criterion of row reduction, the ill-determined pairs are broken up and the matrix becomes well behaved and less subject to roundoff error.

When the procedure steps through the wavelengths a second time, the same selection criterion naturally seeks out the wavelength best correlated with the desired constituent. As each new wavelength is added, the solution to the linear equation uses all of the collected background wavelengths to calculate a background-corrected calibration. Because the best correlation with the concentration of the desired constituent is stored, the wavelength which is retained is the one that shows the greatest ability to be background-corrected.

One might initially surmise that it would be better to drop the smallest rather than the largest B value during the wavelength-stepping procedure. However, because bands in the near-infrared portion of the spectrum are strongly overlapped, precise background correction is critical for a successful calibration. Dropping the smallest B value during the row-reduction process would keep only those wavelengths which are most highly correlated with the desired constituent and would fail to provide adequate background correction.

II. EXPERIMENTAL

A set of simulated spectra was used initially to test the row-reduction algorithm. Four series of random numbers were used to simulate the absorbance spectrum of four pseudo-species at 15 pseudo-wavelengths in each spectrum. Ten pseudo-samples were generated by combining randomly selected amounts of each of the four pseudo-species. The spectrum of each sample was then calculated from a strict application of Beer's law, assuming additivity of the absorbances of the sample constituents. After the simulated spectra were computed, various levels of random noise were added to the spectral and concentration values.

In the first real test of the new algorithm, a set of absorbance data for methyl-red and methyl-orange mixtures, obtained from reference 9, was used to predict solution pH. The data consisted of absorbances obtained at discrete wavelengths ranging from 375 to 575 nm.

In order to compare the new algorithm with those employed earlier, a set of 100 near-infrared diffuse-reflectance spectra of ground wheat samples was obtained from the USDA, Beltsville, MD, and used to predict the percent protein in wheat. Each sample had been assayed for protein by 32 replicate

Kjeldahl determinations. The exact description of the data set has been published elsewhere.[10] The data were used as received with the exception that only every fourth wavelength was considered, for a total of 125 wavelengths. These 125 wavelengths ranged from 1 to 2.6 μm in increments of 12.8 nm. The reported instrumental bandpass was 7 nm and no spectral averaging was used. Fifty samples were used to train the new algorithm and the remaining 50 were used to test it.

Finally, a set of 94 absorbance spectra of synthetic mixtures of benzene, cyclohexane, n-heptane, and iso-octane was used to predict the concentrations of benzene. These spectra were obtained from a Digilab FTS 15C Fourier-Transform infrared spectrometer at a resolution of 8 cm^{-1}. A spectral range of 1.67 to 2.5 μm was considered. Of the 94 measured spectra, 47 samples were used to train the algorithm and 45 samples were used to test it. Two sample spectra were discarded because of verified instrumental error during their acquisition.

III. RESULTS

A. Simulated Spectra. Experiments with simulated spectra simplified the evaluation of the row-reduction algorithm under varying conditions. Several general trends were apparent from these experiments: 1) when no noise was added to the simulated spectra, the algorithm generated an exact solution to Eq. (1) with a SEE of 0; 2) when noise was selectively added, the algorithm consistently chose those wavelengths with the least noise; 3) when additional simulated wavelengths were added but which contained no information (i.e. were not related to sample composition), they

were never chosen when the signal-to-noise ratio of the overall spectrum was greater than 12; 4) when the signal-to-noise of a spectrum was less than 12, the algorithm was less able to distinguish between wavelengths containing information and those containing no information. The probability that an invalid wavelength would be chosen increased as the signal-to-noise ratio decreased. These trends show that the new row-reduction algorithm is viable as long as the signal-to-noise ratio of a spectrum is large enough to make any data reduction worthwhile.

B. Methyl Red / Methyl Orange Spectra. The correlation with pH in mixtures of methyl-orange and methyl-red solution spectra gave statistical correlations ranging from $r = 0.9798$ to $r = 0.9999$, as shown in Table I. These results clearly indicate that the row-reduction algorithm performs well for real solutions where Beer's law is obeyed.

C. Determination of Protein in Wheat. The correlation for protein in wheat is shown in Table II. These results compare well with those obtained by the technique of curve fitting.[11] It should be noted that the number of samples and the number of wavelengths examined at a time in the row-reduction algorithm are necessarily equal because of the fundamental relationship of m independent equations for m independent unknowns in linear algebra. The correlation obtained by the row-reduction method using 7 wavelengths is shown graphically in Fig. 2; the wavelengths and their respective B coefficients (cf. Eq. 1) are listed in Table III.

D. Determination of Benzene in Hydrocarbon Mixtures. The correlation for benzene in hydrocarbons is shown graphically in Fig. 3; the wavelengths used and their respective B coefficients are listed in Table IV. The two

samples plotted as circles in Fig. 3 were known to be in error because of an inadequate instrumental N_2 purge. These latter samples have not been used in calculating the least-squares line, but were retained on the plot to illustrate the possible effect and magnitude of instrumental errors. Although 47 points are plotted in Fig. 3, the precision of the prediction is such that many of the points are not spatially resolvable.

IV. DISCUSSION

A. Computational Efficiency of the Row-Reduction Algorithm. The prediction of protein in wheat shown in Table II and Fig. 2 verifies that the row-reduction algorithm is competitive with other regression techniques as far as standard error of prediction is concerned. There are other considerations, however, which favor row-reduction over multilinear regression. One of these considerations is computation time.

The number of multiplications and divisions required to solve Eq. 1 for a single matrix is equal to:

$$[-N_W^3 + 3(N_S + 1)(N_W)^2 + (3N_S + 4)\, N_W - 6N_S]/6 \qquad (11)$$

where N_W is the number of wavelengths in Eq. 1 and N_S is the number of samples in the training set. In turn, the total number of matrices which must be solved to obtain a calibration via row reduction is approximately the total number of wavelengths to be considered (N_λ) times the number of passes through the wavelength set. Because the number of passes is usually 2, the number of matrices to be solved is ordinarily $2N_\lambda$. Multiplying the number of multiplications and divisions per matrix (Eq. 11) by the number of matrices ($2N_\lambda$)

gives the total number of computations (N_R):

$$N_R = [N_\lambda][-N_W^3 + 3(N_S + 1)(N_W^2) + (3N_S + 4)\, N_W - 6N_S]/3 \qquad (12)$$

The number of multiplications and divisions necessary to obtain a NIRA calibration by "all possible pairs" or "all possible triplets" stepwise regression can be deduced by a two-part computation. The first part is the calculation of the cross terms:

$$\#\ \text{Cross Terms} = \frac{(N_\lambda)(N_\lambda + 1)}{2} \qquad (13)$$

Each of these terms is composed of N_S multiplications so the total number of computations for determining the cross terms is:

$$\#\ \text{Cross Term Multiplications} = \frac{(N_\lambda)(N_\lambda + 1)(N_S)}{2} \qquad (14)$$

The second part of the regression calculation is the inversion of matrices. Each i-by-i matrix inversion requires i^3 multiplications and divisions. The number of matrices to be inverted by the all-possible-pairs stepwise regression is

$$\frac{(N_\lambda)(N_\lambda - 1)}{2} + 2(N_W - 2)(N_\lambda) \qquad (15)$$

The corresponding number for the all-possible-triples stepwise regression is

$$\frac{(N_\lambda)(N_\lambda - 1)(N_\lambda - 2)}{6} + 2(N_w - 3)N_\lambda \qquad (16)$$

where both Eq. 15 and 16 assume one checkback per wavelength addition.

From Eqs. 15 and 16 and the number of multiplications and divisions required to invert each matrix, Eqs. 17 and 18 can be obtained.

$$\text{No. of calculations in all-possible-pairs stepwise regression} = 4(N_\lambda)(N_\lambda - 1) + 2\sum_{i=3}^{N_w} i^3(N_\lambda) \qquad (17)$$

$$\text{No. of calculations in all-possible-triples stepwise regression} = (9/2)(N_\lambda)(N_\lambda - 1)(N_\lambda - 2) + 2\sum_{i=4}^{N_w} i^3(N_\lambda) \qquad (18)$$

An examination of Eq. 12 and Eqs. 14 plus 17 or Eqs. 14 plus 18 gives a semiquantitative basis of comparision of the row-reduction and regression methods. This comparison is tabulated in Table V. It can be observed that row-reduction becomes much more efficient as $N_\lambda \gg N_s$.

B. Other Advantages of Row Reduction. Row reduction has several advantages over regression other than computational efficiency. These advantages include an increased immunity to baseline drift and to overfitting.

If spectral baseline drift occurs, all wavelengths shift up or down together. Therefore, the offset caused by these shifts can be avoided if the B coefficients of Eq. (1) add to zero. The set of typical B coefficients shown in Table III, calculated by row reduction, add very nearly to zero. This feature is inherent in the row-reduction algorithm and avoids the problem of forcing the sum of the regression coefficients to zero. Regression techniques do not inherently possess this feature.

Overfitting occurs when the solution to Eq. (1) reflects trends in the training sample set that are not present in the prediction set. Row-reduction helps reduce the likelihood of overfitting through its use of only the most orthogonal samples to determine the B coefficients. This selection prevents averaging or diluting the uniqueness of individual samples and forces the prediction to be valid for the most unusual samples in the training set, not for the most "typical" samples.

Finally, row-reduction allows an a priori test for overfitting even if the samples whose constituents are to be predicted have not been chemically determined (i.e., are not part of the training or prediction sets). In particular, if the spectrum of a new sample (at the wavelengths used in Eq. (1)) cannot be formed by some combination of the spectra of the samples used to solve Eq. (1), that sample cannot be accurately predicted. This method for detecting the presence of overfitting will be discussed in a subsequent paper.

V. CONCLUSION

The new row-reduction algorithm appears to be a valid technique for finding the correlation between chemical composition and the absorbance or reflectance spectra for spectrally and chemically complex samples. Row reduction has the advantages of computational ease and increased resistance to spectral errors compared to regression methods. Finally, row reduction is conceptually more facile than a multilinear regression, a feature which should aid future research in and interpretation of the NIRA technique.

ACKNOWLEDGEMENTS

This research was supported in part by the Technicon Instrument Corporation, by the Office of Naval Research, and by the National Science Foundation.

REFERENCES CITED

1. J. R. Hart, K. H. Norris, and C. Golumbic, Cereal Chemistry 39, 94 (1962).
2. D. R. Massie and K. H. Norris, Trans. ASAE 8, 598 (1965).
3. I. Ben-Gera and K. H. Norris, Israel J. Agric. Res. 18, 177 (1968).
4. I. Ben-Gera and K. H. Norris, J. Food Sci. 33, 64 (1968).
5. C. A. Watson, Anal. Chem. 49, 865A (1977).
6. N. R. Draper and H. Smith, Applied Regression Analysis, (John Wiley & Sons, New York, 1966).
7. Howard Mark, Technicon Instrument Corp., private communication (1982).
8. B. Kolman, Elementary Linear Algebra, 2nd ed., (Macmillan, New York, 1977).
9. R. M. Wallace, J. Phys. Chem. 64, 899 (1960).
10. P. C. Williams, K. H. Norris, R. L. Johnsen, K. Standing, R. Fricioni, D. MacAffrey, and R. Mercier, Cereal Foods World 23, 544 (1978).
11. W. R. Hruschka and K. H. Norris, Appl. Spectrosc. 36, 261 (1982).

TABLE I. Correlation with pH in Mixtures of Methyl-Red and Methyl-Orange Solutions.

Chemical System	Number of Samples	Number of Wavelengths Retained	Correlation Coefficient (r value)
Methyl Red	4	2	0.9798
Methyl Orange	4	2	0.9999
Mixture	5	2	0.9934

TABLE II. Prediction of Percent Protein in Wheat Using the Row-Reduction Algorithm

Number of Wavelengths Retained for Prediction* by Row-Reduction Algorithm	Number of Samples Used - Both Methods	Reliability of Predicted Percent Protein			
		Row-Reduction		Reference 11	
		SEE	SEP	SEE	SEP
2	2	1.36	1.15	2.28	2.16
3	3	0.40	0.46	2.30	2.30
4	4	0.38	0.38	0.243	0.30
5	5	0.36	0.36	0.24	0.30
6	6	0.31	0.35	0.243	0.30
7	7	0.27	0.26	0.14	0.15

*Reference 11 uses 300 wavelengths for each prediction

TABLE III. Seven-wavelength Correlation for Protein in Wheat
(See also Fig. 2)

WAVELENGTH (μm)	MULTIPLIER (B value)
1.73	900.3
1.74	-967.6
1.86	5.1
1.97	34.2
2.15	1.3
2.17	42.9
2.52	-18.0

TABLE IV. Eight-wavelength Correlation for Benzene in Hydrocarbons. (See also Fig. 3)

WAVELENGTH (μm)	MULTIPLIER (B value)
2.011	0.00073
2.023	0.00971
2.164	0.11889
2.168	-0.26828
2.171	0.29825
2.175	-0.24175
2.179	0.12348
2.189	-0.03678

TABLE V. Number of Computations for Finding the Best 5 and 6-Wavelength Correlations by Row-Reduction and Regression Methods.*

Number of Wavelengths to Search	Number of Samples	Regression methods		Row Reduction
		All possible Pairs stepwise	All possible Triples stepwise	
19	10	20 K	43 K	12 K
19	25	21 K	44 K	32 K
19	50	27 K	51 K	64 K
140	10	297 K	12 M	90 K
140	25	446 K	12 M	232 K
140	50	692 K	13 M	470 K
700	10	5.0 M	515 M	449 K
700	25	8.7 M	519 M	1.1 M
700	50	14 M	525 M	2.4 M

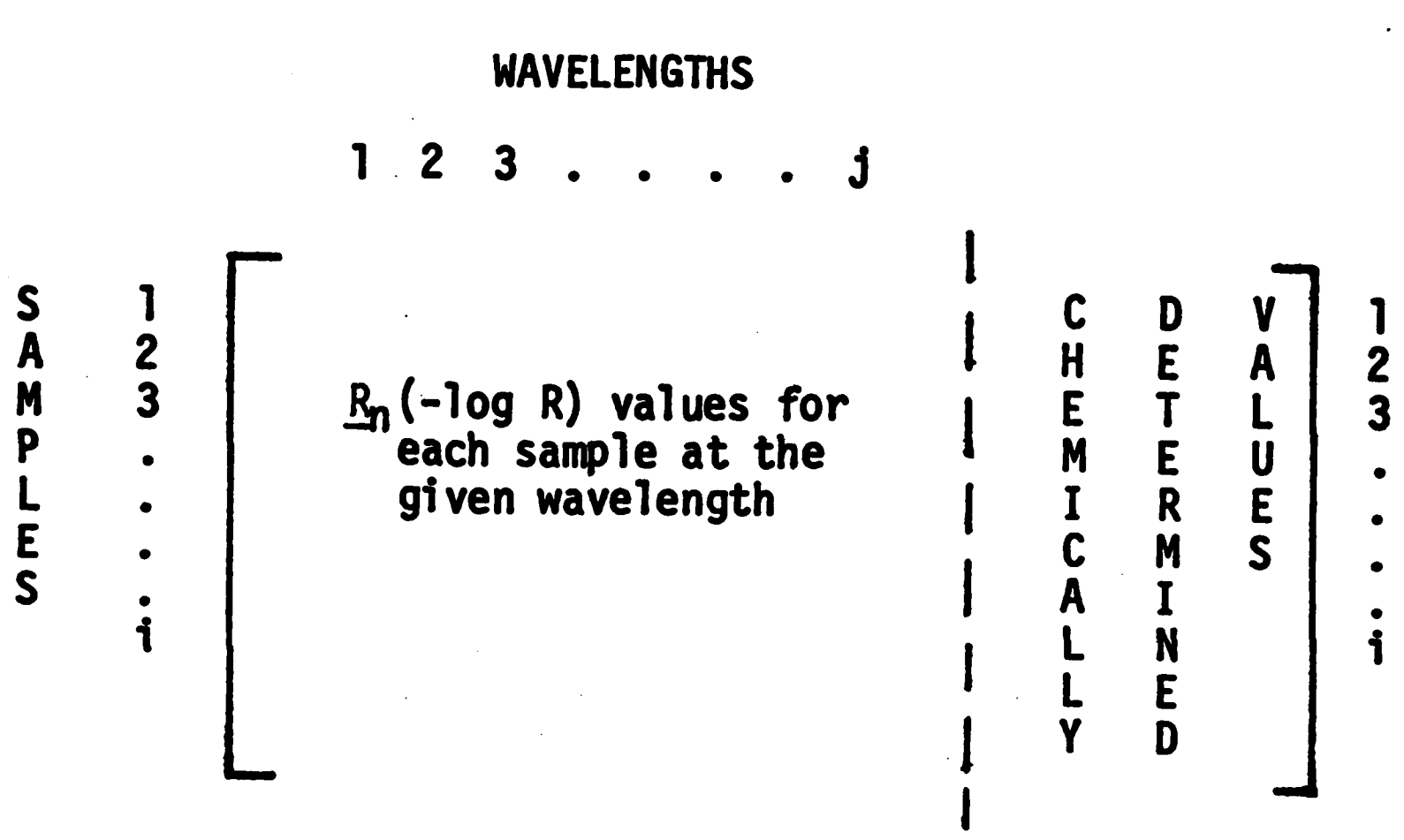
* K $\equiv$ x 10^3; M $\equiv$ x 10^6

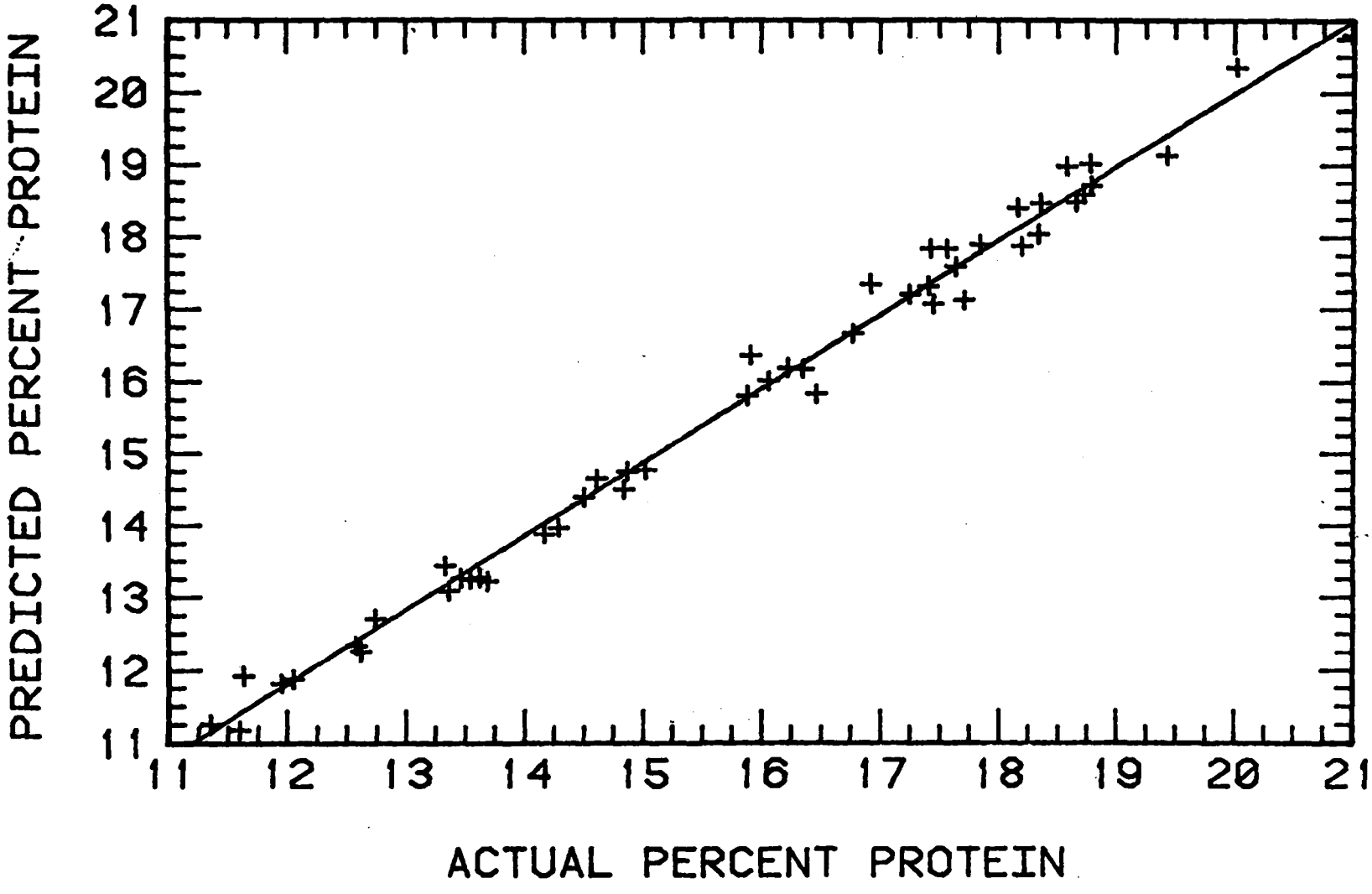
FIGURE CAPTIONS

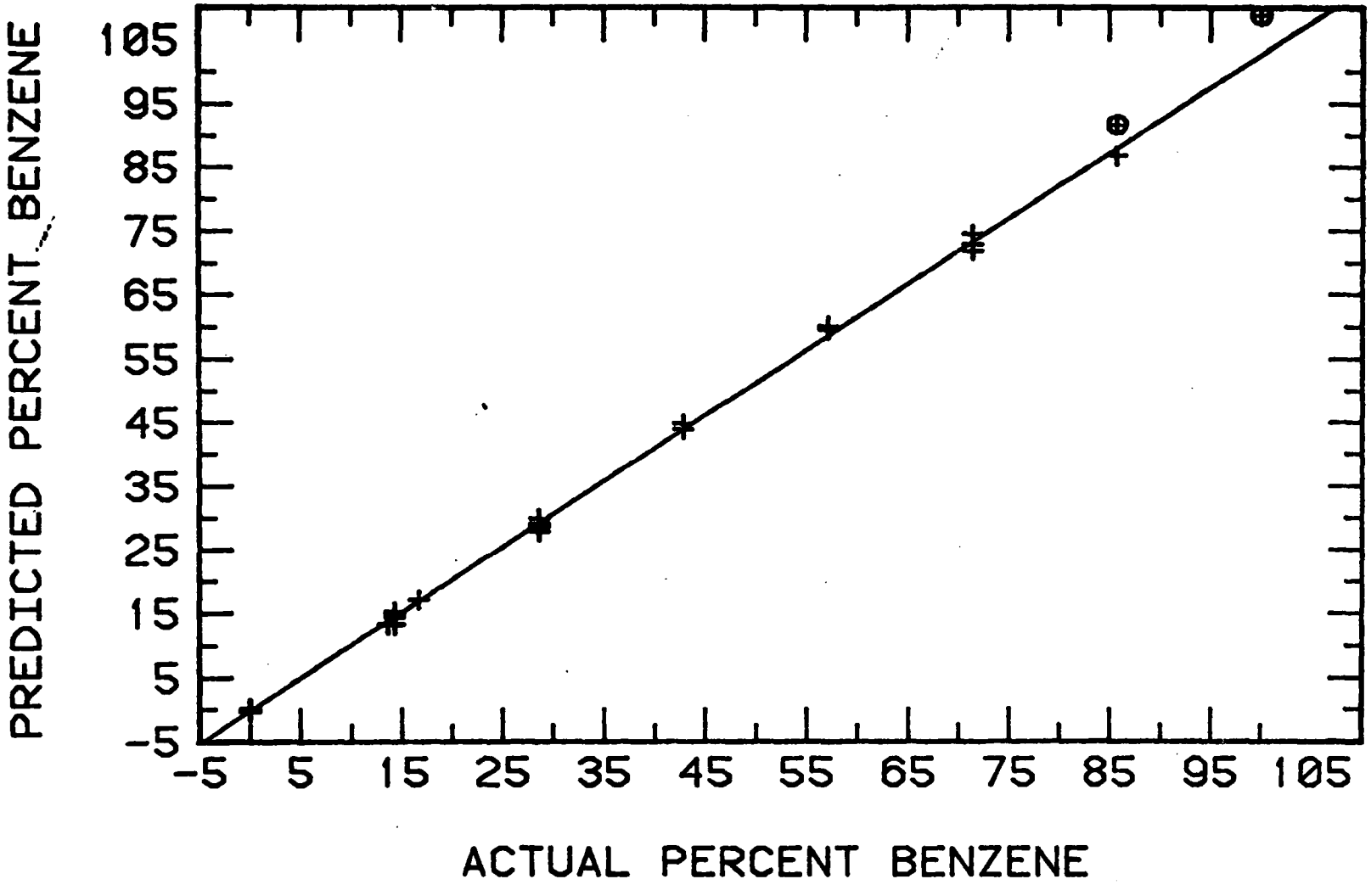
Figure 1. Data configuration of the row-reduction matrix.

Figure 2. Predicted vs. actual percent protein in wheat using new row-reduction algorithm. Fifty samples were predicted using seven wavelengths with a SEP of 0.26% protein.

Figure 3. Predicted vs. actual percent benzene in hydrocarbons. Crosses represent valid data points. Circles represent data points with instrumental errors. Forty-five samples were predicted using eight wavelengths with a SEP of 1.01% benzene.


WAVELENGTHS
1 2 3 . . . . j
SAMPLES
1
2
3
. . .
i
R_n (-log R) values for each sample at the given wavelength
CHEMICALLY DETERMINED VALUES
1
2
3
. . .
i


PREDICTED PERCENT PROTEIN
21
20
19
18
17
16
15
14
13
12
11
11
12
13
14
15
16
17
18
19
20
21
ACTUAL PERCENT PROTEIN


PREDICTED PERCENT BENZENE
105
95
85
75
65
55
45
35
25
15
5
-5
-5
5
15
25
35
45
55
65
75
85
95
105
ACTUAL PERCENT BENZENE

TECHNICAL REPORT DISTRIBUTION LIST, 051C

	No. Copies
Dr. M. B. Denton Department of Chemistry University of Arizona Tucson, Arizona 85721	1
Dr. R. A. Osteryoung Department of Chemistry State University of New York at Buffalo Buffalo, New York 14214	1
Dr. J. Osteryoung Department of Chemistry State University of New York Buffalo, New York 14214	1
Dr. B. R. Kowalski Department of Chemistry University of Washington Seattle, Washington 98105	1
Dr. S. P. Perone Department of Chemistry Purdue University Lafayette, Indiana 47907	1
Dr. D. L. Venezky Naval Research Laboratory Code 6130 Washington, D.C. 20375	1
Dr. H. Freiser Department of Chemistry University of Arizona Tucson, Arizona 85721	
Dr. H. Chernoff Department of Mathematics Massachusetts Institute of Technology Cambridge, Massachusetts 02139	1
Dr. A. Zirino Naval Undersea Center San Diego, California 92132	1

	No. Copi
Dr. L. Jarris Code 6100 Naval Research Laboratory Washington, D.C. 20375	1
Dr. John Duffin, Code 62 Dn United States Naval Postgraduate School Monterey, California 93940	1
Dr. G. M. Hieftje Department of Chemistry Indiana University Bloomington, Indiana 47401	1
Dr. Victor L. Rehn Naval Weapons Center Code 3813 China Lake, California 93555	1
Dr. Christie G. Enke Michigan State University Department of Chemistry East Lansing, Michigan 48824	1
Dr. Kent Eisentraut, MBT Air Force Materials Laboratory Wright-Patterson AFB, Ohio 45433	1
Walter G. Cox, Code 3632 Naval Underwater Systems Center Building 148 Newport, Rhode Island 02840	1
Professor Isiah M. Warner Department of Chemistry Emory University Atlanta, Georgia 30322	
Professor George H. Morrison Department of Chemistry Cornell University Ithaca, New York 14853	1

TECHNICAL REPORT DISTRIBUTION LIST, 051C

	No. Copies
Professor J. Janata Department of Bioengineering University of Utah Salt Lake City, Utah 84112	1
Dr. Carl Heller Naval Weapons Center China Lake, California 93555	1
Dr. Denton Elliott AFOSR/NC Bolling AFB Washington, D.C. 20362	
Dr. J. Decorpo NAVSEA-05R14 Washington, D.C. 20362	
Dr. B. E. Spielvogel Inorganic and Analytical Branch P. O. Box 12211 Research Triangle Park, NC 27709	
Dr. Charles Anderson Analytical Chemistry Division Athens Environmental Lab. College Station Road Athens, Georgia 30613	
Dr. Samual P. Perone L-326 LLNL Box 808 Livermore, California 94550	
Dr. B. E. Douda Chemical Sciences Branch Code 4052 Naval Weapons Support Center Crane, Indiana 47522	
Ms. Ann De Witt Material Science Department 160 Fieldcrest Avenue Raritan Center Edison, New Jersey 08818	

472:GAN:716-4
94/GEN

TECHNICAL REPORT DISTRIBUTION LIST, GEN

	No. Copies
Office of Naval Research Attn: Code 413 800 North Quincy Street Arlington, Virginia 22217	2
ONR Pasadena Detachment Attn: Dr. R. J. Marcus 1030 East Green Street Pasadena, California 91106	1
Commander, Naval Air Systems Command Attn: Code 310C (H. Rosenwasser) Department of the Navy Washington, D.C. 20360	1
Defense Technical Information Center Building 5, Cameron Station Alexandria, Virginia 22314	12
Dr. Fred Saalfeld Chemistry Division, Code 6100 Naval Research Laboratory Washington, D.C. 20375	1
U.S. Army Research Office Attn: CRD-AA-IP P. O. Box 12211 Research Triangle Park, N.C. 27709	1
Mr. Vincent Schaper DTNSRDC Code 2803 Annapolis, Maryland 21402	1
Naval Ocean Systems Center Attn: Dr. S. Yamamoto Marine Sciences Division San Diego, California 91232	1
Naval Ocean Systems Center Attn: Mr. Joe McCartney San Diego, California 92152	1
Naval Weapons Center Attn: Dr. A. B. Amster, Chemistry Division China Lake, California 93555	1
Naval Civil Engineering Laboratory Attn: Dr. R. W. Drisko Port Hueneme, California 93401	1
Dean William Tolles Naval Postgraduate School Monterey, California 93940	1
Scientific Advisor Commandant of the Marine Corps (Code RD-1) Washington, D.C. 20380	1
Naval Ship Research and Development Center Attn: Dr. G. Bosmajian, Applied Chemistry Division Annapolis, Maryland 21401	1
Mr. John Boyle Materials Branch Naval Ship Engineering Center Philadelphia, Pennsylvania 19112	1
Mr. A. M. Anzalone Administrative Librarian PLASTEC/ARRADCOM Bldg 3401 Dover, New Jersey 07801	1

END

FILMED

3-83

DTIC